Durham University

Durham Research Online

Deposited in DRO:

01 April 2010

Version of attached file:

Published Version

Peer-review status of attached file:

Peer-reviewed

Citation for published item:

Byrne, D. and Yang, K. (2008) 'Conceptual statistical problems in exploring the relationship among volume, outcome context in relation to the organisation of secondary tertiary health provision : an issue of causal inference in non-experimental research.', Radical statistics., 96 .

Further information on publisher's website:

http://www.radstats.org.uk/no096/index.htm

Publisher's copyright statement:

Additional information:

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in DRO
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the full DRO policy for further details.

Durham University Library, Stockton Road, Durham DH1 3LY, United Kingdom
Tel : +44 (0)191 334 3042 | Fax : +44 (0)191 334 2971
https://dro.dur.ac.uk

Conceptual and Statistical Problems in Exploring the Relationship among Volume, Outcome and Context in Relation to the Organization of Secondary and Tertiary Health Provision - an issue of causal inference in non-experimental research.

David Byrne and Keming Yang

1. Introduction

In contemporary societies where politics seems no longer to be a matter of class interests or profound ideological differences – in what Crouch (2000) has called post-democracy – many crucial social issues are presented in technical terms with the resolution of those issues dependent on 'scientific' evidence. Much, indeed most, of such evidence is statistical in form and therefore depends on the actual data deployed in arguments and on the procedures used to interpret that data. This paper addresses the way in which 'evidence' about the health outcomes has been interpreted and the way in which those interpretations have been deployed in relation to the 'restructuring' of central elements of health provision in localities in England. We have to set this in context. The 'restructuring' of 'local health economies,' i.e. of systems for the delivery of hospital and related services, is not occurring in a political vacuum. On the contrary 'New Labour' has actively engaged in the introduction of for profit provision into what was historically an overwhelmingly publicly provided not for profit hospital system. However, this agenda can generally only be pursued if some existing hospitals or crucial elements of them – especially maternity and accident and emergency services – are closed. Such closures are generally unpopular with local people

and politicians in post-democratic England, reflecting on the experience of Kidderminster where a local Doctor defeated a seating Labour minister campaigning on the closure of the local hospital, are scared stiff of the political consequences. In particular New Labour MPs fear that enough votes will be drawn away from them to cause them to lose their comfortable and well remunerated parliamentary seats.

Minsters, civil servants and senior health managers realize that nobody trusts them to put it in the vernacular – as far as they could throw them. However, they believe that people trust clinicians. So they need to set up 'clinical cases' for reconfiguration – to get physicians to say that reconfiguration is necessary in order to deliver better health care and better health outcomes. The Blairite think tank, the Institute for Public Policy Research, fired an early shot here (see Farrington-Douglas and Brooks 2007). Byrne and Ruane (2007) responded on behalf of Keep our NHS Public. The Dept of Health wheeled out its senior clinicians (see Alberti 2007, Boyle 2007) and the Darzi report on restructuring health in London (2007) takes the same line. A crucial argument here is that bigger is better, that larger units are better for patients. The 'evidence' for this is drawn from the literature on the relationships between volume and outcome for specific procedures. This paper explores the conceptual and statistical problems which exist in relation to understanding the linkages between volumes and outcomes, much of which is recognized by academic authors but which has been systematically ignored by policy makers. What we have here is a case of policy based evidence, not of evidence based policy – of the misuse of statistical studies, many of which are themselves deeply flawed, to promote political agendas.

2. The Debate in the Literature

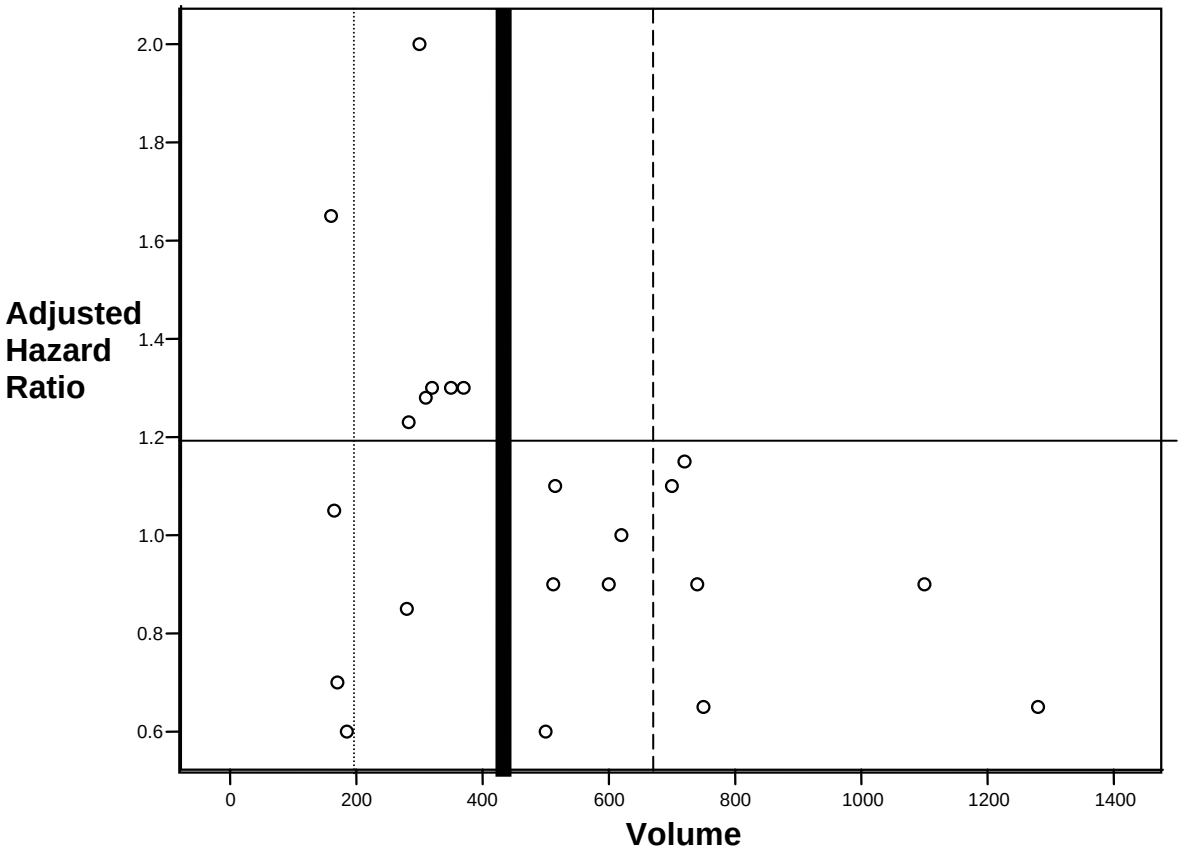
There is an extensive literature dating from Luft's (1979) work of nearly thirty years ago exploring the relationship between volume and outcome in relation to a series of discrete procedures (primarily but not exclusively surgical) with outcome explored in terms both of organization unit volumes and, for surgery, in relation to the volume of work of individual clinicians. These studies explore the statistical associations between volume and outcome, where volume could be the amount of procedures carried out by either of both of health units or individual surgeons, and outcome the

results for patients usually in terms of survival for a period after the procedure. In the literature the evidence on the relationship between volume and outcome has been contested. It is generally agreed that much of the evidence produced before the mid 1990s, which was addressed in a major systematic review by Sowden et al. (1997) was flawed particularly by inadequacies in addressing the issue of differentiating among complexity of cases in relation to specific procedures. Simply put it took no account of how ill people were prior to surgery. So, units which treated only simple cases could do very well, whilst those taking on difficult cases could do badly. Recently Murray and Teasdale (2005) have conducted a literature review of subsequent publications (they specify that this is not a systematic review) which draws heavily on two systematic reviews conducted by Halm et al (2002) and Gandjour et al (2003). Their review is in general and marked contrast to Sowden et al., and comes to the conclusion that there is at least 'sparse' evidence of a causal relationship between volume and positive outcomes whilst agreeing that: 'The relevance of the observed volume/outcome relationships to health service planning depends crucially on how one interprets the underlying mechanisms which generate the associations.' (Ibid 10-11)

3. The Statistical Issues

We agree absolutely with Sowden et al. above but want to propose that the clinical literature as a whole has not understood causal mechanisms properly and that this fundamental misunderstanding means that any simplistic conception of relationships between volume and context is always going to be misconceived. It is important to say that some of the commentators on this issue are plainly aware of this issue and note it with real concern. At best what is observed are associations and the old dictum that correlation is not cause applies here with particular force. And that 'at best' matters because the general approach to the treatment of the data, represented in summary form by the columns which specify 'Number with Significant Volume/Outcome Association' in the Tables which constitute Appendices II and III of Murray and Teasdale's review is fundamentally flawed. Let us deal first with this statistical issue, partly because a lot of studies do something which is really wrong here, but primarily because consideration of this issue provides us with way into thinking about the real mechanisms which generate health outcomes.

Volume is always an interval scale variable (interval scale because we don't count fractions of a procedure) which can quite properly be treated as if it was a continuous variable since it has full ratio properties. In the great majority of these studies outcomes can also be treated as a continuous variable. The only exception are the rather limited number of studies in which data is available about individual outcomes treated usually as a binary dependent variable – often living or dead – in a logistic regression. There are problems even with this sort of use of the general linear model in exploring what are almost certainly non-linear and emergent systems but at least we do not have a measurement crime being committed in those studies. If we have two continuous variables, here volume and some index derived from aggregation of outcome data, then the first essential exploration of any relationship is through the generation of a scatter plot to show the shape of that relationship. Generally in minority of the studies I have examined in which this has been done the plot takes a form like that in Figure One (see Shahian et al. 2003 and Durairaj et al. 2005 for examples. This is modelled on Durairaj.)



Categorisation of volume

If we just look at this graph, we can see that the differences in outcome between high and low volume units depends very much on where we make the cut to differentiate them. Cut at the thick black line i.e. at a volume of 410 and we will have a strong significant difference with bigger volume units doing better, cut at the dashed line at a volume of 620 and that difference will disappear. It all depends on where the cut is made and that cut is made very differently in different studies. This issue has been noted by both Halm et al. and Shahian et al.

> *Even when a compelling volume-outcome relationship existed, our review revealed wide variations in the definitions of high and low volume for a given topic. This made it difficult to specify evidence-based recommendations about which institutions or physicians are truly high-volume providers suitable for 'selection referral'. For almost every condition or procedure for which at least three studies were identified, the thresholds used to define high and low volume overlapped substantially, that is the definition of high volume on one study was the number used to indicate low volume in another. (Halm et al 2002 516)*

> *The lack of a formal statistical approach to the identification and estimation of the volume change point is one of the most troubling aspects of the volume outcome debate. (Shahian et al 2003 1053)*

There is an extensive literature on the issue of dichotomizing continuous variables. Such dichotomizing can take the form either of putting in a cut point which simply divides a distribution into high and low values or putting in two cut points and comparing the high and low extremes whilst not using the middle section data in the analyses. Both approaches have been used in the volume/outcome debate. Both are wrong – see Royston et al 2006, Streiner 2002. Examples of criticism of this procedure can be found in the literature on psychology, market research, epidemiology and clinical practice. If it is done at all there must be very careful specification of cut points and an explicit justification in relation to the shape of the data distribution.

Issues that arise from the context

Two assumptions are implied in the above discussion. First, volume is assumed to be the cause of outcome. Albeit intuitively sensible, this, however, may not be true. It is possible – we need further qualitative studies to verify the mechanisms – that better outcomes by some clinicians attract more patients to them - outcome causes volume. Outcome and volume can reinforce each other as well, forming a causal loop over time. Taking outcome as the effect (or dependent variable) may make sense from a practical point of view, but that should not be understood as the only possible causal process in reality. For non-experimental studies, the identification of the true mechanism or the true causal process has been the greatest challenge. More importantly, health researchers need to constantly remind themselves that that problem cannot be solved by employing statistical methods. In this particular case, there is no way of verifying the causal relationship between volume and outcome by analyzing the scatter plot or calculating any bivariate correlation coefficient.

Second, the dichotomization or categorization of volume has an understandable practical motivation – once a medical unit is labelled 'high' or 'low', then a series of resources and managerial procedures could follow. Such reasoning assumes that a cutting point does exist and only problem is to find it out. Clearly, this is simplistic and the data – given they were properly collected – may show different scenarios. Here, the idea of 'a cutting point' is established based on statistical significance of the difference of two sample means (or proportions). Most researchers with some statistical training would know that whether a statistical test is significant or not depends on a number of things, including the shape of distributions (usually assumed to be the same), variance, and sample size. Therefore, while the researcher is drawing a line separating 'the high' from 'the low', she is also changing the variance and sample size of each side. In essence, it is not just the volume that determines the significance, some other things are at work as well. Things will become more complicated if we are open to the idea of multiple cutting points, and different ways of grouping the cases will result in different conclusions – some groups may show significant relationship between volume and outcome while others don't. It is clear that we have to look somewhere else for the underlying causes.

It is important that the data distribution must be a full description of all cases. It is absolutely wrong to partition continuous variables on the basis of sample derived data since we cannot know from a sample (unless we have elaborate stratification procedures in place) what is the nature of the actual distribution of the variable in the source population. Even when elaborate stratification procedures were adopted in drawing the sample, there is no guarantee that the shape representing the relationship of the two variables in the sample exactly mimics that in the population. The best scenario is that the two variables under study are used as the stratifying variables in drawing the sample, which is rarely the case. Even when that is the case, we still have the uncertainty brought about by sampling errors. Given the satisfaction of a certain number of conditions, sample data could help us estimate the magnitude and the variation of a single attribute or the relationship of two or more variables, but they are not really good for representing the overall structure embedded in the target population. The variation of selected cases across samples therefore brings an additional source of uncertainty to the process of identifying a cutting point or threshold.

Cutting sample derived continuous measurements introduces a level of imprecision which renders all measures of statistical significance meaningless. We should note that given the cross-sectional nature of almost all volume/outcome studies i.e. they rely on data collected at for a single interval – usually one year, it is reasonable to regard the data as a sample from all time intervals. There is the additional complication that such 'time based' samples are never independent, [especially during such a long period as one year, in which events in the early months may induce following reactions later in the year] with the very important exception in the volume/outcome instance of the likelihood that uncommon events may be stochastic. In other words for small units random and independent differences in outcomes may occur since a very small number of deaths might modify the proportionate outcome substantially. This matters as great deal since as Shahian et al note:

> *The performance of a few exceptionally low volume providers is responsible for the significant results in many studies. If these were considered as a separate aberrant group rather than being forced into a global functional relationship with the remaining hospitals, such studies would likely demonstrate a*

less significant volume-outcome association over the intermediate range of volumes. (2003 1053 – the discussion on this page of this article is an excellent summary of the statistical issues as a whole).

4. Working with non-experimental data

So why do studies use 'statistical tests of difference' so regularly when attempting to describe volume/outcome relationships, given that this is a really bad way of dealing with non-experimental continuous data? In our view this is because of the fetishization of the Randomized Controlled Trial in the clinical literature. RCTs explore the impact of a single categorical variate intervention – usually the double or triple blind delivery of drug/placebo, in relation to an outcome which may be measured at any level. They are of course founded on the notion that the condition being treated has a simple chain of causation which can be broken by a simple intervention – not quite the doctrine of specific aetiology but derived from that doctrine. So we establish simple solutions to simple problems by simple interventions with our knowledge derived from experimental interventions in reality. If we dichotomize volume we can then think of volume as analogous to treatment/placebo and the way to assess difference is by a significance test. So that is what we do.
However, the reality in volume/outcome relationships is that:

1. The data is derived from observation of reality rather than experimental abstraction from reality.
2. There is no single intervention which can be categorized but rather volume is a continuous variable which should not be categorized in a simple minded fashion.
3. Causality does not take the form of a simple chain of events but is evidently complex and contingent.

It is this last which matters the most in assessing evidence here. Again several of the authors in the medical literature, particularly those authors who have reviewed the general literature, have a real sense of this issue.

> *… the volume/outcome literature looks at average effects. Although high volume is associated with good outcome in general, there are low volume hospitals whose outcomes are superior to high volume hospitals and there are high volume surgeons with poor results who work within high volume hospitals. (Murray and Teasdale 2005 9)*

> *Even when a significant association exists, volume does not predict outcome well for individual hospitals or physicians. (Halm et al. 2002 517)*
>
> *Most of these studies have used conventional statistical methods that do not recognize the fact that hospitals or surgeons with similar volumes may have very different outcomes because of systematic differences in processes of care, a phenomenon that exaggerates the true statistical significance of the effect of volume on outcome. (Panageas et al. 2003 658)*

We can see this clearly if we look at the horizontal line in Figure One above. That line has been drawn to partition the outcome variable into two – which is actually legitimate since that is the evident data pattern whereas there is no pattern on volume! However, if we look at the bottom left hand corner of the graph we see low volume providers with good outcomes. Indeed more units with less than 200 cases have good outcomes than have bad outcomes. This is a clear indication that we are dealing with complex causal processes.

The literature which does use at least reasonable statistical techniques to explore the volume/outcome relationship begins to recognize this. For example Birkmeyer et al. (2003) who used logistic regression, with outcome being not an aggregate but the actual micro data for individual patients, found that surgeon volume actually accounted for a large proportion of the apparent effect of hospital volume for many procedures. However, we need to go beyond this to look at hospitals as a whole. Urbach and Baxter (2004) note that there are often stronger relationships between outcomes for surgical procedures and the volume of <u>other</u> surgical procedures than between outcomes for that procedure and the volume of that procedure. So something may be going on in large hospitals – as those authors note having to do with resources and quality improvement practices, which is transferable to smaller units. As Duraijai et al put it:

> *Patient volume, a structural construct, has no direct effect on outcomes by itself and is likely to be a proxy for others structures or processes of care. (2005 1687)*

All the systematic reviews and careful critiques of statistical methods are very cautious about volume/outcome relationships and frequently assert that the studies provide very little in the way

of a basis for health provision re-organization. Bigger means better is a lovely simple slogan but in a complex world it is flat wrong.

Conclusion

So do we give up looking for a relationship between volume and outcome? No. If we look at this issue, particularly for National Health Services where we have good data collection and access to all units in the population which whilst quite large is very far from infinite, then we can use techniques of systematic comparison combining qualitative and quantitative research procedures to explore the multiple configurations which generate outcomes. Indeed with developed clinical data bases we can do this in relation to outcomes for individual patients. Ragin's (1987) technique of Qualitative Comparative Analysis is very well suited to the exploration of complex pathway processes leading to outcomes and Blackman and Byrne and Griffiths and Byrne (current work) are developing the application of this procedure in health management and clinical processes. We ought to investigate what generates good outcomes and develop our health systems so as to achieve them but volume/outcome is at best a clue, not a cause, and should never be asserted as the clinical argument for health service reconfigurations which have very different motivations and origins.

The Guardian runs a regular Saturday column written by a clinician which deals with 'Bad Science'. In fact almost all the examples cited in that column deal with either the absence of proper statistical procedures or the crude misuse of data. The English 'clinical cases' for health reconfigurations are classic examples of bad science in a particular political context. 'Radical Statistics' has generally tended to concentrate on statistics as data, as measurements. The volume – outcome – clinical case example shows that we have to pay attention to statistical methods and their use, as well as to data. In relation to 'evidence' in political arguments, this is going to be ever more important across the whole range of public policy and at all levels of society. There are real fights to be had here nationally and locally. Bad Science must not be used as a cover for the privatization of our public health service.

References

Alberti, G. 2007 *Emergency Access: Clinical Case for Change* London: Dept of Health

Birkmeyer, J.D., Stukel, T.A., Siewers, M.P.H., Goodney, P. P., Wennherg, D.E. and Lee-Lucas, F. 2003 'Surgeon Volume and Operative Mortality in the United States' *N Engl Jnl Med* 349 2117-27

Boyle, R. 2007 *Mending Hearts and Brains: Clinical Case for Change* London: Dept of Health

Byrne, D.S. and Ruane, S. 2007 *The Case for Hospital Reconfiguration: Not Proven* London: Keep Our NHS Public http://www.nhscampaign.org.uk/uploads///documents/Reconfiguration%20Not%20Proven.pdf

Crouch, C. 2000 *Coping with PostDemocracy* London: Fabian Society

Darzi, A. 2007 *London's Health Services: A Framework for Action* Health Care for London NHS

Durairaj, L., Torner, J.C., Chrischilles, E.A., Sarrazin, M.S.V., Yankey, J. and Rosenthal, G.E. 2005 'Hospital Volume-Outcome Relationships Among Medical Admissions to ICUs' *Chest* 128 1682-1689

Farrington-Douglas, J. and Brooks, R. 2007 *The Future Hospital: The Progressive Case for Change* London: Institute for Public Policy Research

Gandjour, A. Bannenberg, A., Lauterbach K.W. 2003 'Threshold volumes associated with higher survival in health care – a systematic review' *Med Care* 41 1129-1141

Halm, E.A., Lee, C., Chassin, M.R. 2002 'Is Volume related to Outcome in Health Care?' *Ann Int Med* 137 511-520

Luft, H.S., Bunker, J.P., Enthoven, A.C. 1979 'Should operations be regionalized? The empirical relation between surgical volume and mortality' *N Engl Jnl Med* 301 1364-9

Murray, G.D. and Teasdale, G.M. 2005 'The Relationship between Volume and Health Outcomes' *Report to Volume/Outcome Sub-*

Group – Advisory Group to National Framework for Service Change NHS Scotland

Panageas, K.S., Schrag, D., Riedel, M.A., Bach, P.B., and Begg, C.B. 'The Effect of Clustering of Outcomes on the Association of Procedure Volume and Surgical Outcomes' *Ann Intern Med* 139 658-665

Ragin, C. 1987 *The Comparative Method* London: University of California Press

Roystan, P., Altman, D.G. and Sauerbrei, W. 2006 'Dichotomizing continuous predictors in multiple regression: a bad idea' *Statistics in Medicine* 25 1 127-41

Shahian, D.M. and Normand, S.T. 2003 'The Volume-Outcome Relationship: From Luft to Leapfrog' *Ann Thorac Surg* 75 1048-58

Sowden AJ, Grilli R and Rice N. 1997 *The relationship between hospital volume and quality of health outcomes.* CRD report 8, part 1. York: Centre for Reviews and Dissemination.

Streiner, D.L. 2002 'Breaking Up is Hard to Do: The Heartbreak of Dichotomizing Continuous Data' *Can Jnl Psychiatry* 47 68-74

Urbach, D.R. and Baxter, N.N. 2004 'Does it matter what a hospital is "high volume" for?' *BMJ* 328 737-740

Professor David Byrne
Dr Keming Yang
School of Applied Social Sciences
Contact: dave.byrne@durham.ac.uk